PO2RDF: Representation of Real-world Data for Precision Oncology Using Resource Description Framework

Yiqing ZHAO

Mayo Clinic Minnesota <https://orcid.org/0000-0003-2874-8136>

Anastasios Dimou

Mayo Clinic: Mayo Clinic Minnesota

Feichen Shen

Mayo Clinic: Mayo Clinic Minnesota

Nansu Zong

Mayo Clinic: Mayo Clinic Minnesota

Jaime I. Davila

Mayo Clinic: Mayo Clinic Minnesota

Hongfang Liu

Mayo Clinic: Mayo Clinic Minnesota

Chen Wang (✉ wang.chen@mayo.edu)

Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN. <https://orcid.org/0000-0003-2638-3081>

Technical advance

Keywords: Resource Description Framework, Precision Oncology, Electronic Health Records, Real-world Evidence

Posted Date: February 11th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-219654/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Next-generation sequencing provides comprehensive information about individuals' genetic makeup and is commonplace in precision oncology practice. Due to the heterogeneity of individual patient's disease conditions and treatment journeys, not all targeted therapies were initiated despite actionable mutations. To better understand and support the clinical decision-making process in precision oncology, there is a need to examine real-world associations of patients' genetic information and treatment choice.

Methods: To fill the gap of insufficient use of real-world data (RWD) in electronic health records (EHRs), we generated a single Resource Description Framework (RDF) resource, called PO2RDF (precision oncology to RDF) by integrating information regarding gene, variant, disease, and drug from genetic reports and EHRs.

Results: There are total 2,309,014 triples contained in the PO2RDF. Among them 32,815 triples are related to Gene, 34,695 triples are related to Variant, 8,787 triples are related to Disease, 26,154 triples are related to Drug. We performed one use case analysis to demonstrate the usability of the PO2RDF: we examined real-world associations between EGFR mutations and targeted therapies to confirm existing knowledge and detect off-label use.

Conclusions: In conclusion, our work proposed to use RDF to organize and distribute clinical RWD that is otherwise inaccessible externally. Our work serves as a pilot study that will lead to new clinical applications and could ultimately stimulate progress in the field of precision oncology.

1. Background

Advancement in next-generation sequencing technologies and lowered testing costs have contributed to a much wider embracement of Precision Oncology[1] in oncology clinical practice. The potential of Precision Oncology is to enable oncologist practitioners to make better clinical decisions, by incorporating individual cancer patients' genomic information and clinical characteristics. The anticipation of Precision Oncology is to improve the selection of targeted therapies, avoid side effects from ineffective or toxic therapies, and therefore reduce healthcare costs while improving patient outcomes[2–5].

With increasing needs for Precision Oncology knowledge and evidence, specialized knowledgebases such as OncoKB[6], CIViC[7] and other more general pharmacogenomics or Precision Medicine knowledgebases include PharmGKB[8] and ClinVar[9] were established to curate comprehensive scientific evidence on genes, mutations, drugs, their combined effects on diseases or phenotypes. OncoKB annotates the oncogenic effects and clinical significance of somatic variants[6]. To date, it curated 5,293 unique mutations in 628 cancer-associated genes and 54 tumor types with 92 associated treatment options. Levels of evidence were evaluated based on evidence sources that range from US Food and Drug Administration (FDA) labeling, National Comprehensive Cancer Network guidelines, disease-focused

expert group recommendations to scientific literature[6]. OncoKB provides 300 mutation-treatment associations that were considered actionable. CIViC is also an expert-curated knowledgebase for interpretation of clinical relevance of both inherited and somatic variants in tumors[7]. To date, CIViC contains 3,530 curated interpretations of clinical relevance for 3,075 variants affecting 437 genes among which among 2,250 are treatment related. The interpretations were curated from published literature primarily over the last five years, each interpretation was associated with one or two evidence records. While knowledgebases attempt to generate and evaluate evidence based on literature, it is hard to generalize individual findings from the literature. For example, even though CIViC curated 2,250 are treatment-related evidence, only 16 assertions (knowledge generated from available evidence) regarding 9 genes and 13 mutations were confirmed and published.

Due to the heterogeneity of the Precision Oncology patient cohort, sample sizes for patients in Precision Oncology literature is often small, and patient characteristics are unique. Therefore, it's especially difficult to conduct large-scale clinical trial research or synthesize evidence into knowledge based on different Precision Oncology studies. In a real-world setting, not all targeted therapies were initiated despite the existence of actionable mutations. With the increasing accessibility of digital real-world data (RWD), using RWD to generate real-world evidence (RWE) can be an alternative, low-cost option to bridge the evidentiary gap between clinical research and practice. RWD is defined as data that are routinely generated or collected in the course of health care delivery[10]. Under the 21st Century Cures Act, the FDA developed a program to evaluate the use of RWE to support approval of new indications for approved drugs or to satisfy long-term drug safety surveillance[11]. However, there are challenges for the effective utilization of RWD. One of the challenges includes a limited number of patients with comparable clinical characteristics within one institution. Therefore, it is desirable to increase the interoperability of RWD so that data can be integrated across multiple institutions. Large-scale consortiums such as The Cancer Genome Atlas (TCGA)[12] and Genomics Evidence Neoplasia Information Exchange (GENIE)[13] aim to create centralized databases to address this issue. Another approach to enhance interoperability is by using Wide Web Consortium (W3C) technologies, which provide a set of widely established standards[14]. The Resource Description Framework (RDF) is a recent W3C recommended semantic web tool designed to standardize the definition and use of metadata[15]. It provides a data model that can be extended to address sophisticated ontology representation techniques[15]. In this paper, we described our work that focused on increasing the interoperability of RWD by proposing a novel framework to capture RWD and then represent it using RDF. Based on RWD collected from an institutional oncology cohort, we generated PO2RDF that can potentially be used for downstream analysis. We demonstrated one potential use case of PO2RDF: an examination of real-world associations between EGFR mutation and the prescription of targeted therapies.

2. Methods

In this study, we generated an integrative and standardized data resource for RWD of Precision Oncology via multiple steps, 1) we semi-automatically collected RWD that belongs to key elements (e.g. gene, variant, disease, drug) in a previously proposed precision oncology knowledge model from EHRs; 2) we

normalized the collected data using for further data integration; 3) we integrated collected data using a schema by Genetic Testing Ontology (GTO)[16], which captures the semantic meaning and semantic relations in the collected data; 4) we generated PO2RDF using D2RQ[17]. The workflow performed in this study is shown in Fig. 1

Oncology Cohort

Our cohort includes a total of 2,593 patients with Foundation Medicine tumor mutation test. Foundation Medicine offers three different types of tumor panels and covers a range of 709 genes. All patients in the cohort have granted research authorization and are aged above 18. This research project was approved by the Mayo Clinic Institutional Review Board (IRB# 13-009317) and was following the ethical standards of the responsible committee on human experimentation.

Data Retrieval

Based on the institutional oncology cohort, we semi-automatically collected RWD from genetic reports and the electronic health records (EHRs). Patient IDs were linked to integrate data in genetic reports and EHRs by comparing 1) patient clinic number, 2) first and last name and 3) date of birth. According to our previously proposed precision oncology knowledge model[18], three types of data elements were extracted: “genetic information” (“gene” + “variant”), “disease” and “drug”. Data sources that were used to retrieve three data elements were listed in Table 1. While “genetic information” was extracted from genetic reports only, “disease” and “drug” were retrieved from multiple sources including genetic reports, a unified data platform (UDP), a structured clinical data warehouse of Mayo Clinic[19], and unstructured clinical notes. “Disease” was from both genetic reports and UDP. We extracted only cancer-related diagnosis information. For extraction of “drug” concepts from unstructured clinical notes, we leveraged a dictionary from HemOnc.org[20] that curated comprehensive oncology medication knowledge. Sentences in patients’ clinical notes that mentioned drug concepts were extracted using a natural language processing (NLP) system MedTagger[21]. MedTagger enables a series of NLP processes including dictionary-based concepts indexing, keyword mention lookup, and regular expression matching[22]. Both the drug brand name and chemical name were looked up and were normalized to chemical name.

Table 1
Data Retrieval Sources

	Gene	Variant	Disease	Drug
Genetic Reports	Y	Y	Y	
UDP			Y	Y
Clinical notes				Y

Data Normalization

To facilitate data manipulation and integration, we performed data normalization on RWD extracted from multiple sources. In this study, we mapped “gene”, “variant”, “disease” and “drug” concepts to Unified Medical Language System (UMLS)[23] via the batch process function offered by MetaMap API[24]. The mapping results generated by the MetaMap, include the UMLS preferred terms along with mapping scores. For variants that cannot be mapped to UMLS concepts, we manually normalized variant names to HGVS-nomenclature[25].

Data Integration

We leveraged schema from a previously developed ontology – GTO to integrated the collected RWD. GTO defined seven primary classes namely ‘Diseases’, ‘Gene’, ‘Variant’, ‘Test’, ‘Phenotype’, ‘Risk’ and ‘Drug’ and the relationships among classes[16]. We utilized four of GTO’s primary classes namely ‘Diseases’, ‘Gene’, ‘Variant’ and ‘Drug’ and selected object properties include ‘AssociatedWithGene’ (Domain: ‘Disease’ and Range: ‘Gene’), ‘MayTreatedBy’ (Domain: ‘Disease’ and Range: ‘Drug’), ‘HasContraindicationWith’ (Domain: ‘Drug’ and Range: ‘Disease’), ‘AssociatedWithVariant’ (Domain: ‘Gene’ and Range: ‘Variant’).

We inherited GTO’s data properties especially identifiers that link to external knowledgebases such as Online Mendelian Inheritance in Man (OMIM)[26] and National Drug File Reference Terminology (NDF-RT) [27]. In addition, we added additional identifiers in the data property that link to other precision oncology knowledgebases such as CIViC_Entrez_ID for identifying ‘Gene’ and CIViC_DOID for identifying ‘Disease’ in CIViC. We also incorporated drugs’ brand names (Brand_Name) and categories (Drug_Category) according to HemOnc as additional data properties. We also created a new data class ‘Patient’ to our data schema. The defined data properties for each class, along with some explanation are shown in Table 2.

‘Disease’ and ‘Gene’ relationships were considered valid for diagnosis up to one year before genetic tests. ‘Drug and ‘Gene’ associations (object properties) were considered valid for drug prescriptions up to one year after genetic tests and include targeted therapies only. ‘Disease’ and ‘Drug’ associations (object properties) were considered valid for drug prescriptions after disease diagnosis. For an individual patient, we only count each ‘Disease’ and ‘Drug’ associations once.

Table 2
Description of Data Properties and Related Object Properties

Class	Data Property	Related Object Property
Patient	Patient_ID, Date_of_Birth, Race, Ethnicity, Gender	HasMutGene, HasVariant, HasDisease, TreatedBy
Gene	Gene_Name, UMLS_CUI, OMIM_ID, CIViC_Gene_ID, OncoKB_Gene_ID, PharmGKB_Gene_ID	AssociatedWithGene, AssociatedWithVariant, MayTargetedBy
Variant	Var_Name, UMLS_CUI, ClinVar_ID, dbSNP_ID, CIViC_Var_ID, OncoKB_Var_ID,	AssociatedWithVariant
Disease	Disease_Name, UMLS_CUI, OMIM_ID, CIViC_DOID, OncoKB_Disease_ID, PharmGKB_Disease_ID	AssociatedWithGene, MayTreatedBy, HasContraindicationWith
Drug	Drug_Name, Brand_Name, Drug_Category, UMLS_CUI, NUI (NDF-RT Unique Identifier), CIViC_Drug_ID, OncoKB_Drug_ID, PharmGKB_Drug_ID	MayTreatedBy, HasContraindicationWith, MayTargetedBy

PO2RDF Generation

For the PO2RDF generation, we applied D2RQ, which transforms data in the relational database to RDF. The mapping tool of D2RQ creates a default mapping file by analyzing the schema of an existing database. To map our data with the GTO schema, we manually customized the mapping file accordingly. The data is then published in RDF through the D2RQ server, and can be queried via a D2RQ SPARQL endpoint. We also took an RDF dump from D2RQ into Virtuoso[28] to run federated queries. Figure 2 showed detailed RDF representation of two patients. “Variant” elements were not represented due to space limit.

Use Cases

To demonstrate the usability of PO2RDF, we retrieved triples involving ‘Gene’ and ‘Drug’ from PO2RDF. We then performed association rule analysis[29] to evaluate the significant of real-world associations between mutated gene and selected oncology drugs. Specifically, we examined drugs associated with gene “EGFR”, which are most commonly identified and targeted in lung cancer[30], colorectal cancer[31, 32] and melanoma[33] patients. EGFR inhibitors were initially approved to treat non-small cell lung cancer (NSCLC) and appear to be most effective in patients with adenocarcinoma histology[30]. Even though current FDA drug approved indication for EGRF inhibitors are mostly for NSCLC, it is also used off-label[31–33] for other cancers in real-world setting. Therefore, the results from our association analysis could potentially provide RWE to clinicians and FDA regarding real-world utility of targeted therapies – especially any deviations from guidelines or drug labels.

We calculated the confidence of each {“Drug”, “EGFR”} transaction (Eq. 1). The support of X with respect to a group of transactions T is defined as the proportion of transactions t in the dataset which contains

the item X (Eq. 2). Each individual patient was considered as one transaction (t). Our cohort of 2,593 patients were considered as the total transaction set T.

$$\text{confidence}(X, Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \quad (\text{Eq. 1}); \text{support}(X) = \frac{|\{t \in T; X \in t\}|}{|T|} \quad (\text{Eq. 2})$$

3. Result

Oncology Cohort

We have constructed an oncology cohort of 2,593 (authorized, age > = 18) oncology patients with clinically provided genetic reports. Shown as Fig. 3, this cohort consists of 10 primary types of tumors and is representative of the diversity of patients seen at a dedicated cancer center. As a note, unknown primary cancer cases encompass 10% of the cohort, which indicates the complexity of cases received at Mayo Clinic. In UDP, we were able to retrieve diagnosis codes of 1,193 (46%) patients, among which we were able to identify cancer related diagnosis for 658 patients and 176 received their primary cancer diagnosis at Mayo. This again indicate that heterogeneity of patient population treated at Mayo Clinic – a significant proportion of patients might be referral patients. Thus, combining multiple clinical data sources, especially unstructured clinical notes is crucial to comprehensive RWD capturing. Patient demographic distribution is shown Table 3.

Table 3
Cohort Demographic Distribution

Characteristic	Cohort (n = 2,593)
Average Age at initial diagnosis at Mayo Clinic	58
Average Age at first test	62
Gender (% Female)	51.4%
Race (% White)	88.7%
Ethnicity (% Hispanic)	3.5%

Data Normalization and Integration

To represent PO2RDF in a normal form for further data integration, we mapped individual terms in four classes to UMLS. Table 3 listed the summary of concepts in all four classes. We randomly selected one hundred mapping results for each type of terms and manually reviewed the mapping results. According to our evaluations, there are no incorrect mapping for one hundred ‘Drug’ and ‘Variant’ terms, but there is one incorrect mapping among one hundred ‘Gene’ terms caused due to ambiguity with another disease abbreviation term and two incorrect mappings among one hundred ‘Disease’ terms caused due to substring matching. Despite that ‘Variant’ mappings have been largely accurate, it suffers from huge

missingness mainly due to variations in nomenclature between genetic report and UMLS terminology sources.

Table 4
Statistical Results for Data Collection

	Total number of occurrences	Total number of UMLS-identifiable occurrences	Unique concepts	Unique UMLS-identifiable concepts
Gene	17,100	17,018 (99.5%)	417	415
Variant	16,196	3,158 (19.5%)	5,497	285
Disease	109,030	107,106 (98.2%)	8,449	8,102
Drug	249,995	249,853 (99.9%)	389	368

PO2RDF Generation

There are total 2,309,014 triples contained in the PO2RDF. Among them 32,815 triples are related to Gene, 34,695 triples are related to Variant, 8,787 triples are related to Disease, 26,154 triples are related to Drug. Table 5 include an example SPARQL query and retrieved pertinent information centered on “EGFR”, shown in the “SPARQL Query” column. Specifically, we are searching for related diseases and available targeted drugs, shown in the “Results” column in the Table 5 (for ‘Disease’ and ‘Drug’, only listed top five returned values). An example of data representation of precision oncology evidence from real-world data can be found in Fig. 4. “Variant” elements were not represented due to space limit.

Table 5
SPARQL query to extract EGFR related information

SPARQL Query	Results
<pre>SELECT distinct ?Gene ?property ? hasValue WHERE { ?Gene a po2rdf:Gene. FILTER regex(str(?Gene), "EGFR") ?Gene ? property ?hasValue. }</pre>	<p>Gene_Name: EGFR. UMLS_CUI: C1414313. OMIM_ID: 131550. CIViC_Gene_ID: 1956. OncoKB_Gene_ID: 2. PharmGKB_Gene_ID: PA7360.</p> <p>Disease_Name: 1. Lung cancer, 2. Colorectal cancer, 3. Melanoma, 4. Esophagus adenocarcinoma, 5. Glioma</p> <p>Drugs_Name: 1. Gefitinib, 2. Osimertinib, 3. Afatinib, 4. Erlotinib, 5. Dacomitinib</p> <p>Patient_ID: 3, 15, 21, 44, 65, 73...</p>

Use Case

The result from association analysis was shown in Fig. 5. The top ten EGFR-associated (measured by “confidence”) drugs were listed and they are “gefitinib”, “osimertinib”, “afatinib”, “erlotinib”, “pemetrexed”, “crizotinib”, “cetuximab”, “atezolizumab”, “carboplatin”, and “temozolomide”. The top four drugs are all specific EGFR tyrosine kinase inhibitors (TKIs) and they all have high “confidence” value of association.

Importantly, association rule analysis identified all the EGFR TKIs that are in clinical use in the US. “Confidence” value for “pemetrexed” is significantly lower than the top four reflecting that “pemetrexed” is not a targeted therapy for EGFR mutated cancers. “Pemetrexed” is a cytotoxic chemotherapy drug that can be used to treat mesothelioma and non-small cell lung cancer. “Crizotinib” is also not EGFR targeted therapy. Rather, it is effective in NSCLC driven by activating genomic alterations in “MET”, “ALK” and “ROS1”. Interestingly, although the confidence value for crizotinib and pemetrexed is lower than for specific EGFR TKIs, it is still higher than for carboplatin. This observation reflects the use of crizotinib in combination with EGFR TKIs to treat patients with mutant EGFR positive lung cancer that have developed resistance to EGFR inhibition by acquiring a high MET gene copy number. Additionally, pemetrexed is approved for patients with non-squamous but not for squamous NSCLC, a population enriched in EGFR mutations compared to the population of cancer patients who qualify for treatment with carboplatin. “Cetuximab” is an EGFR inhibitory antibody but it does not show high specificity to EGFR mutations[34]. Overall, the order of confidence values mirrors the prevalence of EGFR mutations in the groups of patients with NSCLC who receive the corresponding drugs. Similarly, association analysis for ALK shown in Fig. 5b, correctly assigned much higher confidence values for all TKIs with ALK specificity, namely crizotinib, lorlatinib, alectinib, brigatinib and ceritinib compared to chemotherapy drugs and immune check point inhibitors that are prescribed in an ALK agnostic manner. The confidence value for crizotinib is lower than for the other ALK TKIs, as crizotinib can also be prescribed to patients with NSCLC and activating genomic alterations in MET or ROS1.

4. Discussions And Future Work

In this study, we introduced a novel precision oncology RDF data resource by integrating heterogeneous information of patients from multiple data sources. Potential use of PO2RDF has been demonstrated in the use case, for example, SPARQL query could facilitate retrieval of comprehensive information regarding genetic mutations and treatment choices by searching the PO2RDF and other relevant and linked knowledgebases. Additional data analytics also demonstrated the potential to use information in PO2RDF for treatment recommendation given a mutated gene. In addition to our demonstrated use case, RDF provides a powerful framework for integrating external data sources e.g. knowledgebases, data from other institutions. Through actively feeding new RWD into PO2RDF, PO2RDF can also serve as a data foundation for a learning health system[35, 36] and can ultimately support the development of clinical decision support system (CDSS) in Precision Oncology practices. If adopted by several institutions, PO2RDF could serve as a tool to enhance interoperability and promote data sharing among participating institutions.

However, there are still challenges in the data normalization phase – even though mapping data in class ‘Gene’, ‘Disease’ and ‘Drug’ to UMLS achieved a high performance, mapping data in ‘Variant’ suffered from low coverage (19.5%). There are two reasons that potentially contribute to the low coverage. (1) In UMLS, variant terms mainly come from two sources: OMIM and National Cancer Institute (NCI). While SNVs has a relatively standardized nomenclature, deletion, insertion, loss, duplication and rearrangement are recorded variably in OMIM, NCI and genetic reports. For example, genetic report variant “CDKN2A

deletion exon 1” will be recorded as “CDKN2A, EXON 1-BETA DEL” in OMIM or simply “CDKN2A Gene Deletion” in NCI. Therefore, it is difficult to extract through regular expression without further normalization. In future work, tools that normalize variant nomenclature to UMLS can be developed to address this unmet need. (2) Both OMIM and NCI have limited records of variant. For example, most frameshift and splice site mutations are not documented in them. A great percentage of fusion cannot be found or can only be mapped partially: “CD74-ROS1 fusion” in genetic reports can only be mapped to “ROS1 Fusion Positive”. Therefore, incorporating of more comprehensive variant knowledgebases such as ClinVar[9] and COSMIC[37] into UMLS is desirable. We also propose to use a structured data entry system supported by clinical terminology in clinical setting for genetic information documentation. This could save time for data input, encourage documentation of genetic information and ensure high quality data capture.

In the future, we plan to apply advance graph mining[38] technologies such as node2vec[39] to discover hidden patterns within the PO2RDF network, which could potentially provide insights to drug repurposing. We also plan to expand data properties by adding temporal information to each data elements. With temporal information, we will be able to make less biased associations between data elements and discover any dynamic pattern changes in the network that may be reflective of disease progression or practice change due to regulatory changes. Moreover, we will incorporate survival data into our PO2RDF, which was currently in the process of data validation. Once survival data is validated, we can provide more valuable insights regarding comparative effectiveness of different drugs.

5. Conclusion

In conclusion, our work proposed to use RDF to organize and distribute clinical RWD that is otherwise inaccessible externally. Our work serves as a pilot study that will lead to new clinical applications and could ultimately stimulate progress in the field of precision oncology.

List Of Abbreviations

Resource Description Framework	RDF
Precision oncology to RDF	PO2RDF
Food and Drug Administration	FDA
Real-world data	RWD
Real-world evidence	RWE
The Cancer Genome Atlas	TCGA
Genomics Evidence Neoplasia Information Exchange	GENIE
Wide Web Consortium	W3C
Genetic Testing Ontology	GTO
Electronic health records	EHRs
Unified data platform	UDP
Natural language processing	NLP
Unified Medical Language System	UMLS
Online Mendelian Inheritance in Man	OMIM
National Drug File Reference Terminology	NDF-RT
Non-small cell lung cancer	NSCLC
Tyrosine kinase inhibitors	TKIs
Clinical decision support system	CDSS
National Cancer Institute	NCI

Declarations

Ethics approval and consent to participate

Research authorizations are obtained in a written format that are compliant with Mayo institutional policy. This project is approved by Mayo Clinic IRB number 20-001474 and 15-003408.

Consent for publication

Not applicable.

Availability of data and material

The data used in this study cannot be shared because of the patient health information included in the texts.

Competing interests

The author(s) declare(s) that they have no competing interests.

Funding

Not applicable.

Authors' Contributions

YZ, HL, CW originated the study. YZ performed NLP analyses, developed the rule-based and machine learning systems and wrote the first draft of the manuscript. SJW and EG evaluated individual clinical cases and provided insights on clinical utilization of PARP inhibitor among patients with *BRCA1/2* mutation. All authors discussed the results and revised the manuscript. All of the authors have read and approved the final manuscript.

Acknowledgements

This research is supported by Genentech Research Fund in Individualized Medicine.

The authors also thank the other investigators from a collaborative team of researchers and information specialists from Department of Health Sciences Research at Mayo Clinic.

References

1. Schwartzberg L, Kim ES, Liu D, Schrag D: **Precision oncology: who, how, what, when, and when not?** *American Society of Clinical Oncology Educational Book* 2017, **37**:160-169.
2. Chantrill LA, Nagrial AM, Watson C, Johns AL, Martyn-Smith M, Simpson S, Mead S, Jones MD, Samra JS, Gill AJ: **Precision medicine for advanced pancreas cancer: the individualized molecular pancreatic cancer therapy (IMPaCT) trial.** *Clinical cancer research* 2015, **21**(9):2029-2037.
3. Evans WE, Relling MV: **Moving towards individualized medicine with pharmacogenomics.** *Nature* 2004, **429**(6990):464-468.
4. Krynetskiy E, McDonnell P: **Building individualized medicine: prevention of adverse reactions to warfarin therapy.** *Journal of Pharmacology and Experimental Therapeutics* 2007, **322**(2):427-434.
5. Ma Q, Lu AY: **Pharmacogenetics, pharmacogenomics, and individualized medicine.** *Pharmacological reviews* 2011, **63**(2):437-459.
6. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH: **OncKB: a precision oncology knowledge base.** *JCO precision oncology* 2017, **1**:1-16.
7. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L: **CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer.** *Nature genetics* 2017, **49**(2):170.

8. Barbarino JM, Whirl-Carrillo M, Altman RB, Klein TE: **PharmGKB: a worldwide resource for pharmacogenomic information.** *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2018, **10**(4):e1417.
9. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W: **ClinVar: improving access to variant interpretations and supporting evidence.** *Nucleic acids research* 2018, **46**(D1):D1062-D1067.
10. Jarow JP, LaVange L, Woodcock J: **Multidimensional evidence generation and FDA regulatory decision making: defining and using “real-world” data.** *Jama* 2017, **318**(8):703-704.
11. Corrigan-Curay J, Sacks L, Woodcock J: **Real-world evidence and real-world data for evaluating drug safety and effectiveness.** *Jama* 2018, **320**(9):867-868.
12. Tomczak K, Czerwińska P, Wiznerowicz M: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.** *Contemporary oncology* 2015, **19**(1A):A68.
13. Consortium APG: **AACR Project GENIE: powering precision medicine through an international consortium.** *Cancer discovery* 2017, **7**(8):818-831.
14. Signore O: **W3C Technologies: a Key for Interoperability.** *Journal of Computer Resource Management (a Publication of the Computer Measurement Group, Inc)* 2003(110):19-40.
15. Decker S, Melnik S, Van Harmelen F, Fensel D, Klein M, Broekstra J, Erdmann M, Horrocks I: **The semantic web: The roles of XML and RDF.** *IEEE Internet computing* 2000, **4**(5):63-73.
16. Li P, Liu H, Zhu Q: **Scientific Evidence Based Genetic Testing Ontology Development towards Individualized Medicine.** *Journal of Translational Medicine & Epidemiology* 2015.
17. Bizer C, Seaborne A: **D2RQ-treating non-RDF databases as virtual RDF graphs.** In: *Proceedings of the 3rd international semantic web conference (ISWC2004): 2004.* Proceedings of ISWC2004.
18. Zhao Y, Yu H, Fu S, Shen F, Davila JI, Liu H, Wang C: **Data-driven Sublanguage Analysis for Cancer Genomics Knowledge Modeling: Applications in Mining Oncological Genetics Information from Patient’s Genetic Reports.** *AMIA Summits on Translational Science Proceedings* 2020, **2020**:221.
19. Kaggal VC, Elayavilli RK, Mehrabi S, Pankratz JJ, Sohn S, Wang Y, Li D, Rastegar MM, Murphy SP, Ross JL: **Toward a learning health-care system—knowledge delivery at the point of care empowered by big data and NLP.** *Biomedical informatics insights* 2016, **8**:BII. S37977.
20. **HemOnc.org - A Free Hematology/Oncology Reference**
21. Liu H, Bielinski SJ, Sohn S, Murphy S, Waghlikar KB, Jonnalagadda SR, Ravikumar K, Wu ST, Kullo IJ, Chute CG: **An information extraction framework for cohort identification using electronic health records.** *AMIA Summits on Translational Science Proceedings* 2013, **2013**:149.
22. Torii M, Waghlikar K, Liu H: **Using machine learning for concept extraction on clinical documents from multiple data sources.** *Journal of the American Medical Informatics Association* 2011, **18**(5):580-587.
23. Bodenreider O: **The unified medical language system (UMLS): integrating biomedical terminology.** *Nucleic acids research* 2004, **32**(suppl_1):D267-D270.

24. Aronson AR: **Metamap: Mapping text to the umls metathesaurus.** *Bethesda, MD: NLM, NIH, DHHS* 2006:1-26.
25. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE: **HGVS recommendations for the description of sequence variants: 2016 update.** *Human mutation* 2016, **37**(6):564-569.
26. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic acids research* 2005, **33**(suppl_1):D514-D517.
27. Simonaitis L, Schadow G: **Querying the National Drug File Reference Terminology (NDFRT) to assign drugs to decision support categories.** *Studies in health technology and informatics* 2010, **160**(Pt 2):1095-1099.
28. Erling O, Mikhailov I: **Virtuoso: RDF support in a native RDBMS.** In: *Semantic Web Information Management.* Springer; 2010: 501-519.
29. Zhang C, Zhang S: **Association rule mining: models and algorithms:** Springer-Verlag; 2002.
30. Gerber DE: **EGFR inhibition in the treatment of non-small cell lung cancer.** *Drug development research* 2008, **69**(6):359-372.
31. Schütte M, Risch T, Abdavi-Azar N, Boehnke K, Schumacher D, Keil M, Yildirim R, Jandrasits C, Borodina T, Amstislavskiy V: **Molecular dissection of colorectal cancer in pre-clinical models identifies biomarkers predicting sensitivity to EGFR inhibitors.** *Nature communications* 2017, **8**(1):1-19.
32. Chan DLH, Segelov E, Wong RS, Smith A, Herbertson RA, Li BT, Tebbutt N, Price T, Pavlakis N: **Epidermal growth factor receptor (EGFR) inhibitors for metastatic colorectal cancer.** *Cochrane Database of Systematic Reviews* 2017(6).
33. Boone B, Jacobs K, Ferdinande L, Taildeman J, Lambert J, Peeters M, Bracke M, Pauwels P, Brochez L: **EGFR in melanoma: clinical significance and potential therapeutic target.** *Journal of cutaneous pathology* 2011, **38**(6):492-502.
34. Douillard J-Y, Pirker R, O'Byrne KJ, Kerr KM, Störkel S, von Heydebreck A, Grote HJ, Celik I, Shepherd FA: **Relationship between EGFR expression, EGFR mutation status, and the efficacy of chemotherapy plus cetuximab in FLEX study patients with advanced non-small-cell lung cancer.** *Journal of Thoracic Oncology* 2014, **9**(5):717-724.
35. Friedman CP, Wong AK, Blumenthal D: **Achieving a nationwide learning health system.** *Science translational medicine* 2010, **2**(57):57cm29-57cm29.
36. Greene SM, Reid RJ, Larson EB: **Implementing the learning health system: from concept to action.** *Annals of internal medicine* 2012, **157**(3):207-210.
37. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E: **COSMIC: the catalogue of somatic mutations in cancer.** *Nucleic acids research* 2019, **47**(D1):D941-D947.

38. Wang Z, Zhang J, Feng J, Chen Z: **Knowledge graph embedding by translating on hyperplanes**. In: *Aaai: 2014*. Citeseer: 1112-1119.

39. Grover A, Leskovec J: **node2vec: Scalable feature learning for networks**. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining: 2016*. 855-864.

Figures

Figure 1

Workflow of RDF Representation of Real-World Precision Oncology Data.

Figure 2

Example of RDF Representation of Two Patients' Data (Purple square: 'Patient', Red circle: 'Gene', Orange circle: 'Drug', Blue circle: 'Disease').

Figure 3

Distribution of Major Cancer Type in the Institutional Oncology Cohort (N = 2,593).

Figure 4

Visualization of RDF Representation of Real-world Associations Among “Gene” (Purple), “Disease” (Green), and “Drug” (Orange) in Precision Oncology. Node size represents the degree of each node (unique number of co-occurrence). Edge thickness represents the weight of each edge (total counts of co-occurrences).

Figure 5

Association Rule Analysis Results (confidence) Regarding Drug – (a) EGFR and (b) ALK mutation Associations.